CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Function2Form Bridge - Towards synthetic protein holistic performance-prediction
Authors	Yallapragada, V. V. B.;Walker, Sidney P.;Devoy, Ciaran;Buckley, Stephen;Flores, Yensi;Tangney, Mark
Publication date	2019-10-07
Original Citation	Yallapragada, V. V. B., Walker, S. P., Devoy, C., Buckley, S., Flores, Y. and Tangney, M. (2019) 'Function2Form Bridge - Towards synthetic protein holistic performance-prediction', Proteins, 88(3), pp.462-475. https://doi.org/10.1002/prot.25825
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25825 - 10.1002/prot.25825
Rights	© 2019, Wiley Periodicals, Inc. All rights reserved. This is the peer reviewed version of the following article: Yallapragada, V. V. B., Walker, S. P., Devoy, C., Buckley, S., Flores, Y. and Tangney, M. (2019) 'Function2Form Bridge - Towards synthetic protein holistic performance-prediction', Proteins, doi: 10.1002/prot.25825, which has been published in final form at https://doi.org/10.1002/prot.25825. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.
Download date	2026-11-05 12:24:28
Item downloaded from	https://hdl.handle.net/10468/8789

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

RESEARCH ARTICLE

PROTEINS: Structure, Function, and Bioinformatics
DOI10.1002/prot.25825

Function2Form Bridge – Towards synthetic protein holistic performance-prediction

VVB Yallapragada [a,b*], Sidney P Walker [a,b,c,d*], Ciaran Devoy [a,b], Stephen Buckley [a,b], Yensi Flores [a,b,c], Mark Tangney [a,b,c]

[a]Cancer Research@UCC, University College Cork, Cork, Ireland

[b]SynBioCentre, University College Cork, Cork, Ireland

[c]APC Microbiome Ireland, University College Cork, Cork, Ireland

[d]School of Microbiology, University College Cork, Cork, Ireland

*Authors contributed equally

Correspondence:

Mark Tangney PhD MBA

m.tangney@ucc.ie

+353 21 420 5709

Running Title: *Function2Form Bridge*

Accepted Article

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/prot.25825

© 2019 Wiley Periodicals, Inc.
Received: Jun 25, 2019;Revised: Sep 02, 2019;Accepted: Sep 17, 2019

This article is protected by copyright. All rights reserved.

Key words:

Synthetic biology, High throughput *in silico* screening, *in silico* modelling, antibody screening, protein scoring, community-based data reporting, machine learning, *de novo* protein design.

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

Abstract:

Protein engineering and synthetic biology stand to benefit immensely from recent advances in *in silico* tools for structural and functional analyses of proteins. In the context of designing novel proteins, current *in silico* tools inform the user on individual parameters of a query protein, with output scores/metrics unique to each parameter. In reality, proteins feature multiple 'parts'/functions, and modification of a protein aimed at altering a given part, typically has collateral impact on other protein parts. A system for prediction of the combined effect of design parameters on the overall performance of the final protein does not exist.

Function2Form Bridge (F2F-Bridge), attempts to address this by combining the scores of different design parameters pertaining to the protein being analysed into a single easily interpreted output describing overall performance. The strategy comprises 1. A mathematical strategy combining data from a myriad of *in silico* tools into an OP-score (a singular score informing on a user-defined overall performance); 2. The F2F-Plot, a graphical means of informing the wetlab biologist holistically on designed construct suitability in the context of multiple parameters, highlighting scope for improvement.

F2F predictive output was compared with wetlab data from a range of synthetic proteins designed, built and tested for this study. Statistical/machine learning approaches for predicting overall performance, for use alongside the F2F plot, were also examined. Comparisons between wetlab performance and F2F predictions demonstrated close and reliable correlations.

This article is protected by copyright. All rights reserved.

This user-friendly strategy represents a pivotal enabler in increasing accessibility of synthetic protein building and *de novo* protein design.

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

INTRODUCTION

The world of in silico aided protein design Proteins are multi-functional biomolecules that perform, mediate and regulate various fundamental functions of life. Over the last 50 years our understanding of proteins and our ability to engineer them has improved dramatically. While proteins find their applications in various fields, biochemical and medical applications have taken the driving seat commercially. Since 1982, when insulin, the first recombinant protein was produced the biotechnological way, the market value for protein-based therapeutics has seen a significant increase, with the market value of bioengineered protein drugs is expected to reach $336.9 billion by 2025[1,2]. Nature has sampled only a small fraction of the theoretical combinations of amino acids that are accessible to proteins[3] , due to the constraints put in place by evolution. Synthetic protein design represents a vast sea of possible space available to be explored. While multiple industries have the potential to exploit non-natural proteins as components of their products or processes, this potential cannot be fully realised without reliable control over protein design.

Understanding the interplay between structure and function of proteins is pivotal in protein design. Techniques such as NMR spectroscopy and X-Ray Crystallography have revealed the structures of over 100,000 proteins and have aided the establishment of protein databases[4], which in turn have facilitated *in silico* protein structure and function prediction based on the amino acid sequence. The need for a faster and affordable way to predict the putative structure of a protein paved the way for computational protein modelling[5]. *In silico*-aided protein design uses

This article is protected by copyright. All rights reserved.

Accepted Article

computational strategy for designing and building proteins that perform a specific function(s), in a user defined setting. Computational methods for structure modelling, docking and function prediction provided *in silico* alternatives for screening protein sequences, creating variants of a specific design[6] and building new *de novo* structures (Figure 1 *in silico* tools) [7]. As a consequence of increasing computational power, the power of *in silico* protein structure and function prediction, and analysis tools is expanding rapidly.

Figure 1:

The overall performance problem: Multiple related parameters and the myriad of in silico tools

Computational tools are now available to predict protein structure, active sites, chemical properties and interactions with other proteins[8-12]. In some cases, these tools could also be used to redesign existing proteins[13] or even design entirely new proteins, in the rapidly evolving field of *de novo* design[14-17]. Unfortunately, a side effect of these rapid advances is that, it is becoming difficult to bridge the gap between these advances in computational technology and their originally intended wetlab applications by biologists. 'Outsourcing' the required *in silico* activity by end users (wetlab biologists) entirely to 'dry-lab' specialists dramatically reduces the potential of *in silico* modelling, 'User empowerment' is key to translating this potential.

This article is protected by copyright. All rights reserved.

Accepted Article

Designing a protein involves defining an overall function (Box 1) and associating it with a 3D structure which is coded into an amino acid sequence[18]. In most cases, the overall function of a protein is a combination of several individual sub-functions. To achieve the overall function, fusing different sub-function parts (Box1) has been the most popular strategy to date. In recent years, *de novo* protein design has been used to obtain amino acid sequences which fold into a required 3D structure for a defined function. In both of these cases, several test sequences are generated to validate their performance against the defined overall function. *In silico* protein design aims to find a 'best-fit' test sequence for a defined overall function (Figure 2).

Box 1: Definitions and commonly used terms

Overall performance: The degree to which the designed protein would perform the defined **Overall Function**

Sub-function part: A part of a protein that is responsible for a particular sub-function. A Sub-function part may represent a sequence for a specific known protein or a part of a protein which has a defined distinct function.

Design Parameters: Parameters such as hydrophobicity, solubility, structure, active site exposure etc., influence the overall performance of a protein. While these parameters are tools for studying the nature of existing proteins, the same parameters, in protein design, can be used as controllers to dictate the overall performance of a protein. Such parameters that dictate the overall performance of a designed protein are grouped under an umbrella term called '**design parameters**'.

The overall performance prediction problem: The complex relationship between the interconnected design parameters and the overall performance.

This article is protected by copyright. All rights reserved.

Accepted Article

The overall performance of a test sequence is a collective functioning of all the individual sub-function parts in concordance. In most cases, the quality of the functioning of the individual sub-function parts is interdependent. Wetlab validation of each test sequence is time consuming, labour intensive and expensive. Moreover, improving a given sequence by modifying individual subfunction part sequences without a holistic analysis on the overall performance represents an *ad hoc* approach prone to low success rates.

Figure 2:

Computational tools are now available to predict protein structure, active sites, chemical properties and interactions with other proteins. These tools score and inform the quality of the individual design parameter using their respective conventional metrics. While these metrics define the quality of the individual parameter in the design, the combined effect of the design parameters (Box 1) on the overall performance is a question yet to be answered. We termed this as 'the overall performance problem' (Box 1). The overall performance problem is a huge challenge in biology wetlab experimentation but not completely new in a mathematical perspective. Recent advances in machine learning and data analyses have solved similar situations. For example, 'selecting a suitable house in a new city' or 'selecting an electronic device that fits my purposes' and 'selecting a car appropriate to my needs and budget'. All these situations are mathematically similar to an extent. The parameters affect the overall performance and the best fit is chosen based on scoring and ranking the potential fits, based on the user's needs. Although the overall performance problem

This article is protected by copyright. All rights reserved.

Accepted Article

has been addressed in other contexts, in the scientific field of protein design, formulating a mathematical model poses a greater challenge. This is due to (i) the lack of a reporting system for negative results, (ii) time constraints in biological experimentation and (iii) the lack of standardization of measurement units in many areas of biological research.

In this work, we propose F2F-Bridge as a novel mathematical strategy aimed at predicting the overall performance of a synthetic protein. Several test sequences were designed for a defined overall function. The individual scores for all the different design parameters pertaining to each test sequence are condensed into a graphical output. The result is a visual and numerical evaluation of the test sequence. The graphical output (F2F-Plot) and the numerical evaluation (OP-score) together form a novel mathematical strategy (F2F-bridge) that scores, ranks and predicts the overall performance of the given set of test sequences. This method combines user input with *in silico* data to give insights into the predicted overall performance of a test sequence. With the view to eventually developing a robust tool for protein performance prediction, we also examined relationships between *in silico* and laboratory data for test proteins using two different strategies for feature selection and predictive model building: LASSO and regression-based decision trees implemented with the RandomForest algorithm.

This article is protected by copyright. All rights reserved.

Accepted Article

MATERIALS AND METHODS

In silico design of test sequences:

Luminescence proteins: Each construct was designed to have a luminescent domain, a binding domain, a solubility tag and a secretion signal. All parts are linked in all possible permutations using different rigid and flexible linker sequences[19] (Figure 3)**.** Variable heavy and light chain AA sequences from different antibodies were used as the binding domains, from an antibody targeting either cell surface associated epithelial mucin 1 (MUC1; mammalian antigen) or Clumping factor A (ClfA) of *Staphylococcus aureus* (bacterial antigen). Test sequences were designed to bind to their respective target and present luminescence as a readout (bound protein luminescence). Fluorescence proteins (used for further validation) are described in Supplementary Text 3.

Figure 3:

This article is protected by copyright. All rights reserved.

Accepted Article

Data Generation:

The different *in silico* features analysed in relation to the overall performance of the test protein, and how they are generated is outlined in Table 1, with more detailed instructions found in Supplementary Text 1.

Function2Form Bridge:

The data described in the data generation stage were taken as input into the Function2Form function, written in the R programming language. The F2F function takes as input a data frame with all proteins to be screened as rows, and the different *in silico* observations as columns. The first row of the table contains a set of user desired input values. Some of these are based on the benchmarks provided by *in silico* programmes such as RC score and Instability, while in other cases the user can specify the ideal needs for the protein (e.g. hydrophobic and < 30 kDa). The features used in this study are detailed in Table 1, but the F2F-Bridge function is not limited to these and can be easily expanded or condensed depending on individual user requirements.

The test sequences are then screened by the function and the output is a data frame of the test sequences and their respective scores, along with a radar plot for each test sequence, highlighting the differences between each sequence and the input parameters as can be seen in Figure 5. Areas where the candidate protein does not meet the preset requirements will appear outside the coloured

This article is protected by copyright. All rights reserved.

region of the reference values. As well as this visual analysis of the suitability of the protein, a score indicating overall function is provided. The OP score is the grand average of absolute distance between each particular feature of the protein, and the user-specified/program specified reference range. For cases of high throughput *in silico* screening, an additional function allows the user to extract the *n* top scoring test sequences from the overall selection.

F2F R function:

Generating the OP score

(i) Where possible, convert different *in silico* observations to same scale

$$i = \left(\frac{(O - O_{min})}{(O_{max} - O_{min})}\right) * (N_{max} - N_{min}) + N_{min}$$

Where O is the old range and N is the new range, which in the case of the F2F function is always 0-100.

(ii) F2F function then iteratively scores each test sequence supplied in input table

$$OP\ Score\ of\ x = \frac{\Sigma|xi - yi|}{n}$$

Where *x* is the test sequence to be scored, *y* is the set of reference values, *i* refers to the *ith* observation within the *in silico* data table supplied to the algorithm, and *n* is the total number of observations *i.*

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

Generating the F2F plot

Figure 5 shows the graphical output of the F2F function. Each sequence tested is assigned an OP score as discussed above, and a radar plot is generated. For every *in silico* observation provided in the input data, an axis is created on the radar plot. This enables the user to see which specific *in silico* feature or features are making the test sequence unfit for purpose.

Statistical and Machine learning methods:

Two other methods of transforming the *in silico* data into a prediction of overall performance were assessed. The aim was to examine the data for relationships of any kind between the predictive features and the wet lab output with the view to either design a system of weights for the predictive features to improve the F2F plot, or to create a new predictive tool to be used in conjunction with the F2F Bridge.

i) *LASSO* Feature selection and subsequent generation of predictive models can be used in conjunction with the F2F-Bridge programme. LASSO regression puts a constraint on the sum of the absolute values of the model parameters, which must be less than a fixed upper limit. It does this by applying a regularisation process (shrinkage) where it penalizes regression coefficients and shrinks a selection to zero. Variables that still have a non-zero coefficient after the shrinking process are selected to be part of the final model [20] .

This article is protected by copyright. All rights reserved.

Accepted Article

ii) *TREE BASED METHOD* The second method was to use a regression tree as per the random Forest package in R to generate a variable importance plot, again using the *in silico* parameters as input, and wetlab luminescence as the indicator of overall performance. An outline of how random forest works in generating these regression trees is as follows. A predefined number of bootstrapping samples are drawn from the original data. For each of these samples, an "un-pruned" regression tree is grown. Traditionally, the best split at each node to differentiate all predictors would be used, but in this instance the best split is found amongst a random subset of the predictors. Following this, predictions are made by aggregating the predictions of the pre-defined number of trees and taking the average value[21]. The quality of the model was ensured by finding the optimal number of features to randomly sample at each split, and to ensure that enough iterations of the model are run to ensure that the out of bag error has stabilised.

Wetlab validation:

The luminescence test synthetic proteins examined are outlined in Figure 3. Two biological facets were used to assess the effectiveness of the functional prediction strategies – i) binding; ii) secretion. Sub-function parts on the test sequences include: (i) **Active site:** Heavy and light chains of anti-MUC1 antibody (C595) and anti-ClfA antibody were fused with EAAAK (rigid) and GGGGS (flexible) linkers to obtain Monospecific bivalent diabodies and Monovalent ScFVs (monobodies), (ii) **Secretion signal:** Gaussia luciferase's native secretion signal, (iii) **Solubility enhancer:** SUMO tag, (iv) **Reporter:** Truncated version of Gaussia luciferase was used as a

This article is protected by copyright. All rights reserved.

luminescence reporter. (v) **Detection tag:** Flag peptide was used as a detection tag for downstream assays. See Fig 4.

Presence or absence of certain sub-function parts or their design orientation has a significant effect on the overall performance of the protein and should be accounted carefully in the design phase. In our case, over 50 different amino acid sequences were designed against each target. Of these, 8 variants per target were synthesised for testing in the wetlab. These test sequences vary in (a) (+/-) solubility enhancer, (b) (+/-) and positioning of Active site and (c) the type/format of Active site. All these test sequences were tested for their overall performance. Wet lab data were used to validate and improve the results from the F2F-Bridge. An outline of the laboratory workflow can be seen in Figure 4, and a more detailed description on synthesis and build of 'test sequences' can be found in Supplementary Text 3.

Wetlab analysis of fluorescence test sequences is described in Supplementary Text 3.

Data generation from wetlab experiments with luminescence proteins:

Binding assays: 10^8 *Staphylococcus aureus* TCH959 (naturally bearing *clfA)* or 10^6 MCF7 cells (naturally bearing MUC1) were blocked with 5% BSA for 2 h followed by incubation with supernatant containing each test construct. Cells were washed 3 times and resuspended in PBS. Luminescence was measured using Promega GloMax® 96 luminometer. In our case, since bound

Accepted Article

This article is protected by copyright. All rights reserved.

luminescence is the overall function, the luminescence readings corresponding to each test sequence are recorded and used for validating and improving F2F Bridge.

Figure 4:

Statistical analysis: All statistical testing, unless otherwise stated, was performed in the base R environment v3.4.3[22]. The LASSO regression feature selection method was implemented using the Glmnet library v2.0-16[23], and the Random Forest regression tree analysis was performed using the RandomForest library v4.6-14[24]. The radar plot within the F2F-bridge function was implemented with the fmsb library, v0.6.3[25]. Visualisation was carried out using the ggplot2 package,v3.1.1[26].

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

RESULTS

Overall Biological Performance - Bound luminescence

16 test AA sequences were scored using F2F-plot. The result of F2F-plot analysis and associated scores for all test sequences can be seen in Figure 5 and Supplementary Figure 1, and a detailed workflow of the strategy can be found in Supplementary Text 2, additionally the raw data can be found in Supplementary Table 1. It can be seen from the output of the F2F-bridge that the *ClfA Monobody 2* test sequence (represented by the pink shaded region) has an instability index score that is far higher than the user required level (blue shaded region), but there is minimal difference between the different shaded regions across the other axes, which contributes to *ClfA Monobody 2* having a low and therefore good score. The test sequence predicted to have the poorest overall performance, *ClfA Diabody 2*, has levels of solvent accessibility and docking affinity that are much lower than required, as well as an increased instability index. These factors combine to give this test sequence the highest and therefore worst score.

Figure 5:

These sequences were used to generate the corresponding proteins as outlined in the Methods section. Biological performance of these proteins was assessed in the wetlab using binding assays with luminescence as the readout, corresponding to Overall Performance.

This article is protected by copyright. All rights reserved.

Accepted Article

Table 2 shows the results (Wetlab – luminescence units; F2F – OP score) of all test sequences whose biological performance was predicted with F2F-Bridge. The accuracy of F2F prediction was assessed by comparing the F2F-prediction of overall biological performance with the laboratory luminescence data.

Overall, the correlation pattern showed the F2F-Bridge method providing a general guide for how the test sequence can be expected to perform. In a database of this limited size, there was no statistically significant correlation between the OP-score and the 'bound protein luminescence' evident. This was repeated with 'luminescence' as the wet lab output, but again there was no relationship present (Data not shown).

Investigation of alternative methods to complement F2F plot

1) LASSO Regression

'Bound luminescence' as overall performance: The output of Lasso analysis of the test sequences for bound protein luminescence can be seen in Figure 6. In the case of test sequences against MUC1, the features deemed to have the most effect on bound protein luminescence were Docking Affinity, Hydrophobicity, Solvent accessibility and Isoelectric point. However, when these predictive features were input into a linear model, no linear relationship was detected. The same was found when analysing the potential relationship between the test sequences against ClfA and

This article is protected by copyright. All rights reserved.

their eventual bound protein luminescence. In this case, Hydrophobicity and Instability were the features selected, and again, no linear relationship was found.

Despite the LASSO method detecting possible relationships between the predictive features and the test variable (bound protein luminescence), no predictive linear model could be constructed (Table 3). We speculate that there may be too many variables present from a wetlab perspective for us to accurately predict bound luminescence with a database of this size.

'Secreted luminescence' as overall performance: To counter this, we also examined 'luminescence' only resulting from protein secretion as a measure of overall biological performance. As we are just measuring the degree of luminescence of the test sequences, and not their binding to a target, the two groups (antiMUC1 and antiClfA) can also be directly compared, doubling our sample size. Figure 6 uses the LASSO method to investigate the relationship between *in silico* observations and luminescence due to secretion of anti–Clfa test sequences. *In silico* observations implicated in dictating the level of secretion of a test sequence were identified by LASSO. These were then used to generate a linear model to examine the degree to which they explained the luminescence due to secretion of the test sequences. This gave a linear model that explained 84.6% of the variability in luminescence of all test sequences, with an associated p-value of 0.004. This led us to explore the utility of a LASSO dictated linear model as a predictive tool. The luminescence levels predicted by the model were correlated with the experimentally determined luminescence levels. This, and also correlation coefficients of individual test sequence

Accepted Article

This article is protected by copyright. All rights reserved.

groups against their luminescence, are shown in Figure 6, and numerically in Table 4. As would be expected, antiClfA test sequences which are the training set, show stronger correlation (Rho 0.93), but that of the antiMUC1 test sequences was also significant.

2) Alternative methods for prediction of test sequence performance; Random Forest Regression Tree analysis

The same methodology was used for a regression tree implemented within random Forest. The test sequences against ClfA were used as the training data set, and the test sequences against MUC1 test sequences were used as the test set. The regression model derived from the random forest algorithm was able to explain 41% of the variability in the luminescence of the training set data test sequences (Figure 6). The values predicted by random forest for luminescence for the individual proteins were then correlated with their experimentally derived levels of luminescence (Table 5). This method showed a significant correlation between the luminescence values predicted by the random forest algorithm, and the experimental values. Random forest regression analysis was carried out with bound protein luminescence as the wet lab output, but no significant results were found (data not shown).

Figure 6:

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

We postulate that with an increased amount of data available for training the models, this accuracy can only increase. A summary of all the tests performed and their outcomes is shown in Table 6.

Further Validation – Fluorescence Proteins

The F2F plot was further validated on a second dataset of 8 test sequences and resulting laboratory data (fluorescence readings from proteins). In this instance there was no need for the secondary functionality of feature selection and model building based on a subset of the data, as the OP score predicted showed a statistically significant inverse correlation with overall performance of the test sequence. These results are presented in Figure 7 and Table 7, and the accompanying F2F plots, scores and raw data are shown in supplementary Figure 2 and Table 2. In this case, fluorescence was the overall function of the 8 test sequences to be predicted. Unlike the dataset discussed previously, in this case, the F2F-plot predicted scores showed a strong inverse correlation with the overall performance of the proteins, which is the desired result.

Figure 7:

This article is protected by copyright. All rights reserved.

Accepted Article

DISCUSSION

In this study, we have shown how F2F plot could help to visualize the overall performance of a test sequence, particularly if complemented by the statistical methods examined. For each sequence, the OP score predicts the expected efficacy, and the F2F plot provides a graphical overview of predicted strengths and weaknesses. Unlike the pre-existing *in silico* tools that inform the quality of individual design parameter, F2F bridge takes a top down approach on overall performance by predicting the collective influence of all the design parameters on given test sequence. Such a holistic outlook on the overall performance holds a key for informed protein design. F2F bridge could be used either for low throughput design (see Table 8), accounting for the 'pitfalls and merits', in the design corresponding to a particular test sequence, or for high throughput *in silico* screening by comparing and ranking a set of test sequences.

By informing the end user (wet lab biologist) with performance predictions, F2F plot and OP score together aim to ultimately bridge the gap between easily generated, seemingly abstract *in silico* values, and experimental results.

When combined with downstream testing, we have shown that patterns exist in the easily generated *in silico* data that can be used to generate predictive models.

F2F Bridge An unweighted and unsupervised combination of features deemed likely to have an effect on biological performance showed promising results. The F2F-Bridge is able to give an early indication of the expected performance of a test sequence. Given the ease of implementation,

This article is protected by copyright. All rights reserved.

Accepted Article

in comparison with wetlab experiments, any information provided about candidate test sequences prior to synthesis is extremely valuable. As well as the OP-score provided, the accompanying radar plot can also highlight any design aspects of the test sequence that diverge from what is required as per the user input parameters. We expect a considerable improvement in performance of both the F2F-Bridge and associated models, with an expanded dataset, in the meantime further work to refine or build on the method was carried out.

Feature selection driven linear models with LASSO LASSO regression was used to search for patterns in the data that could help predict the level of luminescence due to a test sequence binding to its target. Features that have an effect on bound luminescence were identified. When viewing the plots of normalised Lambda1 vs coefficients, what is important is the point at which the predictive feature enters the model, and the effect it has on the dependent variable. In this case, it was impossible to incorporate them into a statistically significant linear model to predict any of the variability in bound protein luminescence. We speculate that, with a database of this small size, there were too many different processes involved from a laboratory perspective for the strategy to accurately predict the final outcome (steps in cell expression of a given test sequence, protein binding to target, luminescence production). For this reason, and the opportunity to increase the sample size, we also analysed the secreted luminescence data. This was much more successful in terms of improving on the F2F-Bridge.

LASSO regression-based feature selection was used again, to look for patterns between *in silico* observations of the antiClfA test sequences and their experimentally determined secreted

This article is protected by copyright. All rights reserved.

Accepted Article

luminescence [20] . This was more effective. We have shown that a linear model can be predicted using LASSO feature selection that can predict values for luminescence that correlate strongly with experimentally determined values. This functioned retrospectively to find patterns between *in silico* features of antiClfA test sequences and their eventual levels of secreted luminescence, it can also function prospectively. Once the linear relationship between a subset of the *in silico* features and luminescence has been established for antiClfA, this was used to successfully predict luminescence values in a test set of antiMUC1 proteins which correlated with the experimental values. The antiMUC1 data originated from a different experiment, and a different class of proteins to the antiClfA test sequence training set, so the fact that the linear model still has predictive power is extremely encouraging.

Random Forest regression trees The random forest regression tree model, given the same *in silico* features as the LASSO regression, was able to explain ~41.01% of the variability in secreted luminescence within the antiClfA test sequence dataset. The predicted secreted luminescence values generated by the regression tree model significantly correlated with the experimentally derived secreted luminescence values. On a group by group basis, it is extremely encouraging that, as with the LASSO based method previously, a random forest regression model trained on the antiClfA test sequences was then able to predict luminescence for the test set (antiMUC1 test sequences) that correlated significantly with the values derived experimentally [21]. Future work, with an expanded database would involve assessing whether these associations identified with the two above methods become stronger as the database size increases, and also, if this method of

This article is protected by copyright. All rights reserved.

predicting overall biological performance holds true for other tasks, opening up the possibility of the design of an accurate prognostic tool for test sequence performance.

Relevance to the laboratory scientist It is important to frame these results in the context of the difference in cost between *in silico*- and laboratory-based screening. *In silico* screening requires a fraction of the time, money or expertise of laboratory-based screening, so maximising the value of this data can lead to considerable operational savings. Wet lab experimentation is often a prolonged process in biological research. In most cases, the data from wetlab assays has to be processed/filtered to observe the intended correlations between the experimental aims.

We have shown that *in silico* predicted protein attributes can play a significant role in optimising the design and production of protein constructs. With a larger sample size, we expect the aforementioned methods to become more accurate, and therefore call for the establishment of a community wide database for sharing both *in silico* and experimental data so that this can be incorporated into a much larger training data set applicable to a variety of biological functions. Inspiration came, in part, from the SourceTracker algorithm used in metagenomic studies to track possible sources of contamination in HTS studies. This involved establishing a database of known contaminants, used by the algorithm to refine the search for contaminant bacteria[27] . We hope that the establishment of a database of potential test sequences used in other research laboratories, combined with their in silico parameters and in the case of those that are synthesised, their biological output, would similarly improve the accuracy of the predictions of the F2F-Bridge. To expedite the process of database formation we plan to launch a server in the coming months which

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

will take as input an amino acid sequence, and perform all of the necessary calculations, and will also accept one or more measures of overall biological function, in the hope that fellow researchers will submit their published data. Both linear models (selected with lasso regression) and tree based methods (random forest regression trees) can further assist in the prediction of "overall biological performance". With larger datasets, we hope to develop one or more of the following strategies:

i) A method of applying weights to the features in the F2F-Bridge based on the outputs of the two models previously mentioned

ii) Develop a distinct predictive tool based purely on one or both of these methods

iii) Design a protocol suited to large scale projects whereby an initial subset of the test sequences are generated and screen experimentally and *in silico*, by the F2F-Bridge itself, or with a combination of the three methods outlines in this paper. The resulting information could then be fed back into the design protocol to refine this process, increasing the success rate of constructs.

Conclusions The design-build-test-learn approach of synthetic biology stands to benefit immensely from this method. Wetlab assays to test multiple test sequences demand a huge amount of resources, time and human effort. In such a situation, Function2Form becomes an indispensable strategy for a biologist to visualise and improve a given test sequence or to triage potential best performers by scoring and ranking the test sequences. Integrating F2F-Bridge into the 'learn' step aids user empowerment by providing a wetlab biologist with a holistic readout on the overall

This article is protected by copyright. All rights reserved.

performance of the protein. With a community-based data reporting system and larger data sets, the accuracy of the F2F-Bridge could be tuned to Pareto optimality.

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

ACKNOWLEDGMENTS

The authors acknowledge relevant support from Science Foundation Ireland (12/RC/2273 and 15/CDA/3630), the Health Research Board (MRCG2016-25) and Breakthrough Cancer Research.

REFERENCES

1. Research TM. Bioengineered Protein Drugs Market Worth US$336.9 Billion by 2025: Increasing Drugs in Phase III of Clinical Trials to Boost the Market. [Market report]. 2017; https://www.prnewswire.com/news-releases/bioengineered-protein-drugs-market-worth-us3369-billion-by-2025-increasing-drugs-in-phase-iii-of-clinical-trials-to-boost-the-market-636479663.html.
2. Flores Bueso Y, Tangney M. Synthetic Biology in the Driving Seat of the Bioeconomy. *Trends in biotechnology.* 2017;35(5):373-378.
3. Huang P-S, Boyken SE, Baker D. The coming of age of de novo protein design. *Nature.* 2016;537:320.
4. Rose PW, Prlic A, Altunkaya A, et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* 2017;45(D1):D271-D281.
5. Prentiss MC, Hardin C, Eastwood MP, Zong C, Wolynes PG. Protein Structure Prediction: The Next Generation. *Journal of Chemical Theory and Computation.* 2006;2(3):705-716.
6. Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. Thermodynamic prediction of protein neutrality. *Proceedings of the National Academy of Sciences of the United States of America.* 2005;102(3):606.
7. Woolfson DN, Bartlett GJ, Burton AJ, et al. De novo protein design: how do we expand into the universe of possible protein structures? *Current Opinion in Structural Biology.* 2015;33:16-26.
8. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.* 2015;43(W1):W174-181.
9. Yang J, Zhang Y. Protein Structure and Function Prediction Using I-TASSER. *Curr Protoc Bioinformatics.* 2015;52:5 8 1-15.
10. Kappel K, Miao Y, McCammon JA. Accelerated molecular dynamics simulations of ligand binding to a muscarinic G-protein-coupled receptor. *Q Rev Biophys.* 2015;48(4):479-487.
11. Li J, Shaikh SA, Enkavi G, Wen PC, Huang Z, Tajkhorshid E. Transient formation of water-conducting states in membrane transporters. *Proc Natl Acad Sci U S A.* 2013;110(19):7696-7701.

This article is protected by copyright. All rights reserved.

Accepted Article

12. Clark AJ, Tiwary P, Borrelli K, et al. Prediction of Protein-Ligand Binding Poses via a Combination of Induced Fit Docking and Metadynamics Simulations. *J Chem Theory Comput.* 2016;12(6):2990-2998.
13. Wood CW, Heal JW, Thomson AR, et al. ISAMBARD: an open-source computational environment for biomolecular analysis, modelling and design. *Bioinformatics.* 2017;33(19):3043-3050.
14. Mitra P, Shultis D, Brender JR, et al. An evolution-based approach to De Novo protein design and case study on Mycobacterium tuberculosis. *PLoS Comput Biol.* 2013;9(10):e1003298.
15. Bellows ML, Floudas CA. Computational methods for de novo protein design and its applications to the human immunodeficiency virus 1, purine nucleoside phosphorylase, ubiquitin specific protease 7, and histone demethylases. *Curr Drug Targets.* 2010;11(3):264-278.
16. Jiang L, Althoff EA, Clemente FR, et al. De novo computational design of retro-aldol enzymes. *Science.* 2008;319(5868):1387-1391.
17. Karanicolas J, Corn JE, Chen I, et al. A de novo protein binding pair by computational design and directed evolution. *Mol Cell.* 2011;42(2):250-260.
18. Coluzza I. Computational protein design: a review. *J Phys Condens Matter.* 2017;29(14):143001.
19. Chen X, Zaro JL, Shen WC. Fusion protein linkers: property, design and functionality. *Adv Drug Deliv Rev.* 2013;65(10):1357-1369.
20. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society.* 2996;58(1).
21. Liaw A, Wiener M. *Classification and Regression by RandomForest.* Vol 232001.
22. Team RC. R: A language and environment for statistical computing. . *R Foundation for Statistical Computing.* 2017.
23. Jerome Friedman TH, Robert Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software.* 2010;33(1):1-22.
24. M.Wiener ALa. Classification and Regression by randomForest. *R News.* 2002;2(3):18-22.
25. Nakazawa M. *Functions for Medical Statistics Book with some Demographic Data.* 2007.
26. Wickham H. ggplot2: Elegant graphics for data analysis. *Springer-Verlag New York.* 2009.
27. Knights D, Kuczynski J, Charlson ES, et al. Bayesian community-wide culture-independent microbial source tracking. *Nature Methods.* 2011;8:761.
28. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. *Nat Methods.* 2015;12(1):7-8.
29. Alhossary A, Handoko SD, Mu Y, Kwoh CK. Fast, accurate, and reliable molecular docking with QuickVina 2. *Bioinformatics.* 2015;31(13):2214-2216.
30. UCLA. SAVES server. 2019; http://servicesn.mbi.ucla.edu/SAVES/.

This article is protected by copyright. All rights reserved.

31. Fraczkiewicz R, Braun W. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *Journal of Computational Chemistry*. 1998;19(3):319-333.
32. Wilkins MR, Gasteiger E, Bairoch A, et al. Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol.* 1999;112:531-552.

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

Figure legends:

Figure 1: In silico tools for protein modelling and design.

Figure 2: The in silico aided protein design process and the overall performance problem.

Multiple test sequences are generated for a user-defined overall function (The overall function is the collective functioning of all the sub-function parts). Sub-function parts and the test sequence build dictate the 3D structure. Test sequences are subjected to in silico modelling and multiple analyses. F2F bridge provides an objective, single metric and a calculated output to aid the selection of an optimal candidate.

Figure 3: 3D structure of the designed luminescence test sequences examined showing various sub-function parts for a defined overall function.

Figure 4: Workflow of wet lab validation of test sequences.

Figure 5: F2F-bridge output for ClfA test sequences.

This article is protected by copyright. All rights reserved.

Accepted Article

Areas where the candidate protein (pink) does not meet the preset requirements is highlighted by contrasting with the coloured region of the reference values (blue). The plot is generated using a bespoke function written in R, and detailed instructions on its use can be found in supplementary materials, with a link to the github repository containing the code.

Figure 6: Output of F2F Post Hoc Analysis

(A) *shows the graphical output of LASSO features selection for both* ***(i)*** *antiMuc1 and* ***(ii)*** *antiClfA test sequences for predicting* ***bound protein luminescence****. Each coloured line corresponds to a predictive feature used in the F2F-Bridge function. The lines plot the path of the variables coefficient against the L1-norm, of the whole coefficient vector as lambda varies. Both show that Lasso regression analysis was capable of identifying relationships between the examined predictive features, and experimentally determined luminescence.*

(B)(i) *shows the same LASSO based feature selection for antiClfA test sequences, using* ***secreted luminescence*** *as output. In this instance, Isoelectric point enters the model first, and Instability appears to have the most pronounced effect on the test variable. The relationship between these and the experimentally derived luminescence was tested with a multiple linear regression. Both features were found to be significant in the model, which explained 84.61% of the variability in luminescence, with an associated p-value of 0.004. This allowed us to build a predictive model using antiClfA test sequences as the training set.* ***(ii)*** *Shows a correlation plot of luminescence values for test sequences predicted by a LASSO directed linear model, vs experimentally derived secreted luminescence values. The training set (antiClfA test sequences) is coloured blue, and the test set (antiMUC1) is coloured red.*

(C) *summarises the Random Forest regression tree analysis. As with* ***(B)*** *the successful model was trained on the antiClfA data, and tested on the antiMUC1 data.* ***(i)*** *shows mean node purity for each predictive feature. The lower this value, the more important it is to the model. The model, trained on the antiClfA test sequences was able to explain 41% of the variability in*

This article is protected by copyright. All rights reserved.

experimentally determined secreted luminescence. The model was then used to predict secreted luminescence values of the training set (antiClfA) and the test set (antiMUC1), and these predicted values were correlated with experimentally derived luminescence values in figure ***(ii)****. The overall correlation coefficient was 0.87, with an associated p-value of 1e-05, indicating the value of the predictive model generated.*

Figure 7: Correlation plot of OP-Score vs Overall Biological Performance.

Overall biological performance was scored for fluorescence. Correlation test carried out using Spearman's method.

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

Table Legends:

Table 1: Common in silico tools, and the purpose they serve

Table 2: Agreement between experimental results and F2F plot - Luminescence. *In both antiMuc1(A) and Anti-ClfA(B), the binding affinity of each protein is measured by luminescence output. Proteins are ranked by their OP score in both tables, and coloured from green (best performing protein), through yellow, to red (worst performing protein) for both luminescence and OP score. (Lowest number = best OP score.)*

Table 3: *Results of multiple regression analysis of features selected by Lasso regression analysis against experimentally determined luminescence.*

Table 4: *Correlation coefficients and associated p-values of LASSO directed linear model based predictions of test sequence luminescence, vs experimentally derived results for luminescence.*

Table 5: *Results of various predictive models of secreted luminescence generated using RandomForest regression trees, when correlated with the lab-generated values.*

Table 6: Summary of statistical tests performed

Table 7: Agreement between experimental results and F2F plot - Fluorescence.

Proteins are ranked by their OP score, and coloured from green (best performing protein), through yellow, to red (worst performing protein) for both Fluorescence and OP score. (Lowest number = best OP score.)

Table 8: The two associated workflows of the F2F Bridge.

This article is protected by copyright. All rights reserved.

Accepted Article

Design Parameter	Metric and Scale range	Metric description	Effect on overall performance	Generated:
3D structure	C-score 2 : -5	Confidence in the model and folds predicted	Higher confidence reflects more reliable model	I-Tasser [28]
Docking	ΔG kcal/mol	Gibbs free energy released by reaction	Protein-protein interactions at the active site predicted	Autodock Vina* [29]
Quality of model	RC – score	Proportion of amino acids in different regions based on steric hindrance	High agreement with stereochemistry and free energy reflects stability of structure	Saves server [30]
Active site solvent accessibility	0-9	A measure of the exposure of A residue or group of residues	Depending upon the function, the active site could be exposed to the solvent or 'hidden' inside the core	R (Using I-TASSER output)/GETAREA server [31]
Surface Hydrophobicity	-4.5 to +4.5	Each amino acid has a hydrophobicity score between -4.5 and +4.5 as per Kyte Doolittle scale.	Ensuring ideal surface hydrophobicity aids solubility	R (Using I-TASSER output)
Size	kDa	Total weight of the protein	Size forms an important factor if the protein is required to cross/penetrate membranes and biological barriers	ProtParam Hosted by Expasy [32]
Isoelectric point	pH 0 to 14	Point at which molecule carries no net charge	The integrity of the structure of the protein in a setting is influenced by the isoelectric point	ProtParam Hosted by Expasy [32]
Potential active sites	0 to n	Number of potential active sites	Predicting the potential active sites on the designed protein informs on potential off-target effects	Coach Server
Instability	0 to 100	Half life of protein *in vitro*	Gives an indication of the viability of the protein	ProtParam Hosted by Expasy [32]

Table 1: Common in silico tools, and the purpose they serve

This article is protected by copyright. All rights reserved.

Accepted Article

A

WetLab output	F2F-Plot Score	Test Sequence
2788	4.69	Muc1 Monobody 2
96232	5.04	Muc1 Monobody 4
106712	5.73	Muc1 Monobody 3
4914	5.79	Muc1 Monobody 1
2156	6.68	Muc1 Diabody 1
2436	6.9	Muc1 Diabody 3
1597	9.27	Muc1 Diabody 6
30	9.48	Muc1 Diabody 2

B

WetLab output	F2F-Plot Score	Test Sequence
9302	3.62	ClfA Monobody 2
15179	3.8	ClfA Monobody 4
26519	5.29	ClfA Monobody3
90901	5.62	ClfA Monobody 1
45542	6.03	ClfA Diabody 3
9507	6.88	ClfA Diabody 1
209	9.27	ClfA Diabody 6
1110	9.48	ClfA Diabody 2



Table 2: Agreement between experimental results and F2F plot - Luminescence.

In both antiMuc1(A) and Anti-ClfA(B), the binding affinity of each protein is measured by luminescence output. Proteins are ranked by their OP score in both tables, and coloured from green (best performing protein), through yellow, to red (worst performing protein) for both luminescence and OP score. (Lowest number = best OP score.)

This article is protected by copyright. All rights reserved.

Test sequences	P-Value	Adjusted R Squared
antiClfA	0.507	-.06
antiMUC1	0.867	-.6

Table 3: *Results of multiple regression analysis of features selected by Lasso regression analysis against experimentally determined luminescence.*

Accepted Article

This article is protected by copyright. All rights reserved.

	All	antiMuc1	antiClfa
Rho	0.846	0.71	0.93
P value	0.0009	0.04	0.0006

Table 4: *Correlation coefficients and associated p-values of LASSO directed linear model based predictions of test sequence luminescence, vs experimentally derived results for luminescence.*

Accepted Article

This article is protected by copyright. All rights reserved.

	antiClfA	antiMuc1	Total
P value	0.002	0.05	1e-5
Rho	0.92	0.71	0.87

Table 5: *Results of various predictive models of secreted luminescence generated using RandomForest regression trees, when correlated with the lab-generated values.*

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

Summary of methods tested and results:

***In silico* Test Performed**	**Biological Performance Tested**	**Test Sequence**	**Result**
"Blind" F2F Bridge	Binding and Luminescence	antiClfA and antiMUC1	The F2F plot was able to provide a guide for the expected performance of the test sequence when the test sequences were ranked by OP score and by wet lab output, and the accompanying plot was able to inform on how to improve the test sequence. No statistically significant relationship between OP score and wet lab output could be found.
LASSO feature selection and linear model **building**	Binding	antiClfA and antiMUC1	LASSO regression analysis was able to detect discrete patterns in the data, showing Hydrophobicity and Isoelectric point both to have a positive relationship with bound luminescence in antiClfA. In the case of antiMUC1 Docking Affinity and Solvent accessibility were shown to have a positive effect, Isoelectric point and Hydrophobicity a negative one.
Using LASSO regression analysis dictated linear model as a **predictive tool**	Binding	antiClfA and antiMUC1	The models predicted in the above analysis were unable to exaplain any of the variability in the bound luminescence of antiMUC1 or ClfA test sequences.
LASSO feature selection and linear model **building**	Luminescence	antiClfA	LASSO regression analysis was able to detect discrete patterns in the data, a linear regression with solvent accessibility and instability was able to explain 86.4% of the variability in luminescence in the antiClfA samples.

This article is protected by copyright. All rights reserved.

Accepted Article

Using LASSO regression analysis dictated linear model as a **predictive tool**	Luminescence	antiClfA and antiMUC1	The model created in the above test was used to predict luminescence values for both antiClfA and antiMUC1. In both cases these predictions showed strong positive correlations with the experimental luminescence values which were statistically significant.
Random Forest regression tree model **building**	Luminescence	antiClfA	A regression tree implemented with randomForest was able to explain ~41% of the variability in the luminescence of antiClfA test sequences.
Using Random Forest regression tree as a **predictive tool**	Luminescence	antiClfA and antiMUC1	The model created in the above test was used to predict luminescence values for antiClfA and antiMUC1 test sequences. In both cases, these predictions showed strong positive correlations with the experimental luminescence values that were statistically significant.

Table 6: Summary of statistical tests performed

This article is protected by copyright. All rights reserved.

Accepted Article

WetLab Output	F2F-Plot Score	Test Sequence
1.52E+09	4.92	Fluorescence Construct 5
1.05E+09	5.61	Fluorescence Construct 2
3.73E+07	6.43	Fluorescence Construct 6
2.91E+07	6.73	Fluorescence Construct 3
3.06E+08	8.67	Fluorescence Construct 1
2.85E+08	10.68	Fluorescence Construct 8
2.52E+07	12.58	Fluorescence Construct 4
1.88E+07	14.35	Fluorescence Construct 7



Table 7: Agreement between experimental results and F2F plot - Fluorescence.

Proteins are ranked by their OP score, and coloured from green (best performing protein), through yellow, to red (worst performing protein) for both Fluorescence and OP score. (Lowest number = best OP score.)

This article is protected by copyright. All rights reserved.

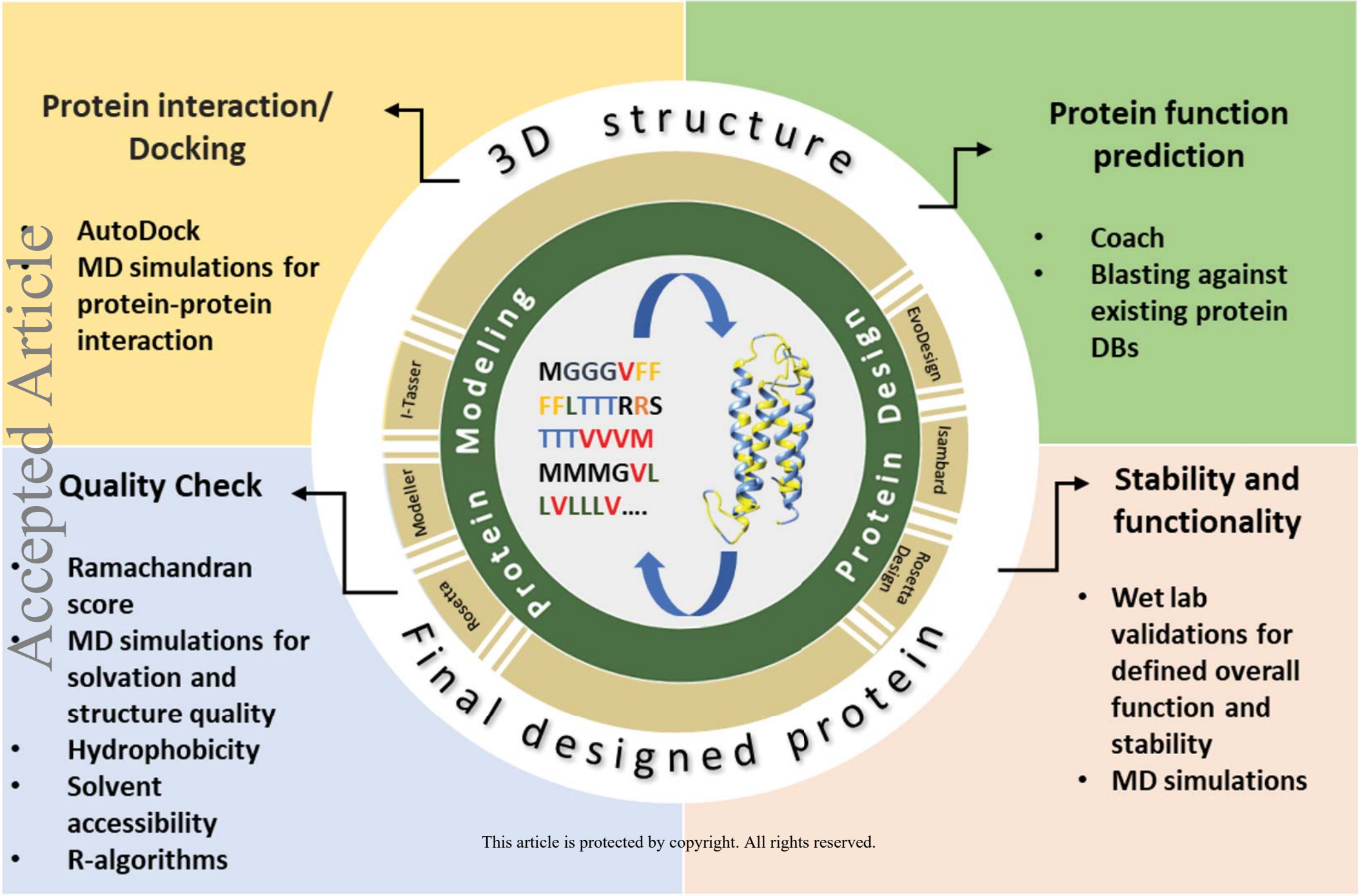
Scale of use	Outcome
High Throughput	A database of test sequences or extant proteins of known sequence can be queried with the F2F-bridge scoring each test sequence and identifying those most suitable.
Low Throughput	On a protein by protein basis the F2F-bridge provides a graphical overview of the relationship between the features of the test sequence and the optimal values specified by the user, informing the user on how to improve the test sequence.

Table 8: The two associated workflows of the F2F Bridge.

Accepted Article

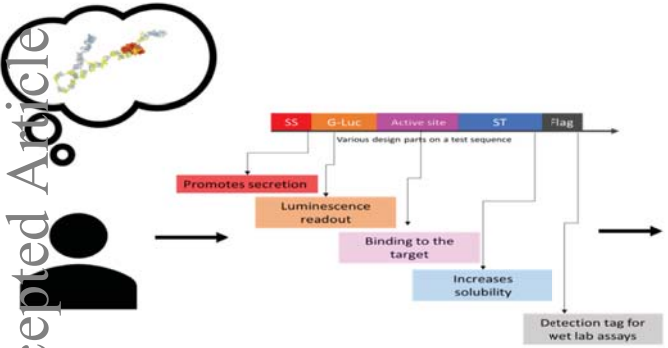
This article is protected by copyright. All rights reserved.

Accepted Article



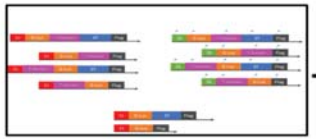
This article is protected by copyright. All rights reserved.

Accepted Article

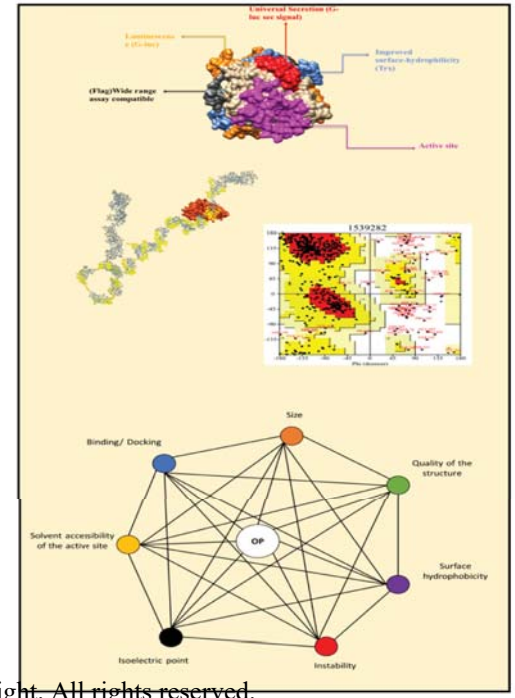


User defined overall function

Sub-function parts



Multiple test sequences

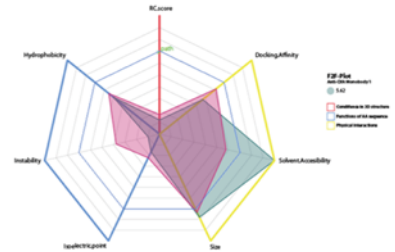


In silico modelling and analysis

Traditional approach

Subjective best guess user decision

F2F Bridge approach

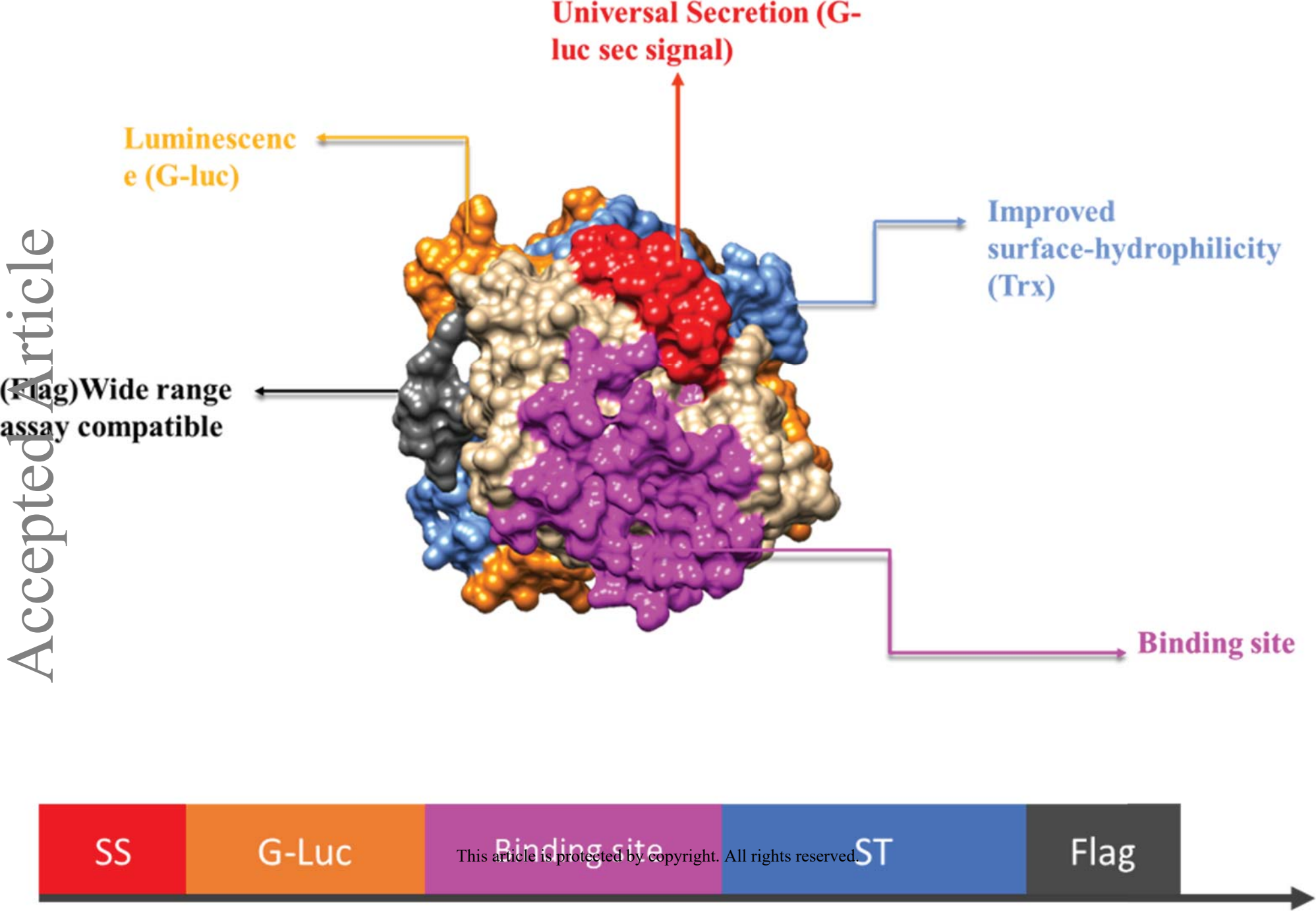


Calculated selection

Selection of optimal candidate

This article is protected by copyright. All rights reserved.

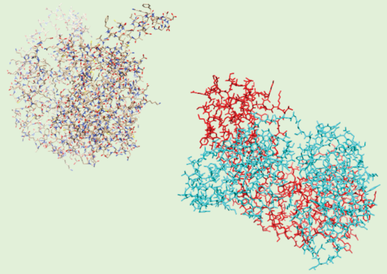
Accepted Article



This article is protected by copyright. All rights reserved.

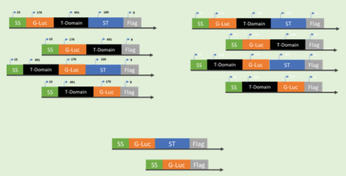
Accepted Article

1



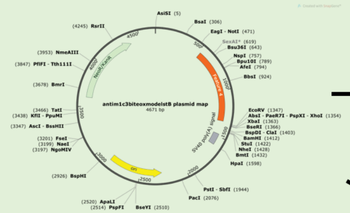
Test sequences

2



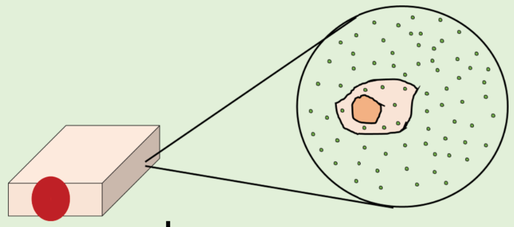
DNA G-Blocks

3



Gibson Assembled into plasmid

4



Transfect CHO cells

Transform bacteria

5

Protein containing liquid medium

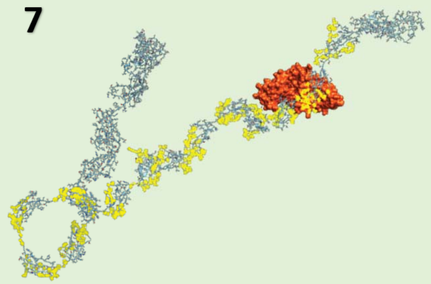


6



Luminescence or Fluorescence

7



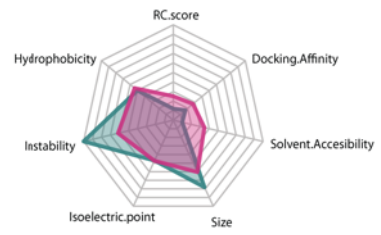
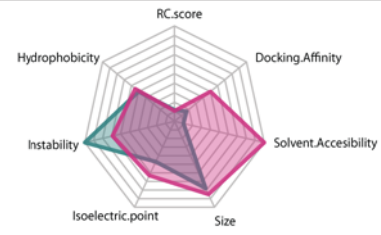
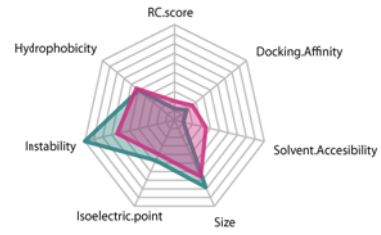
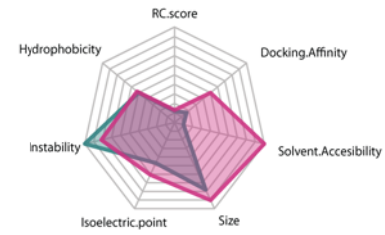
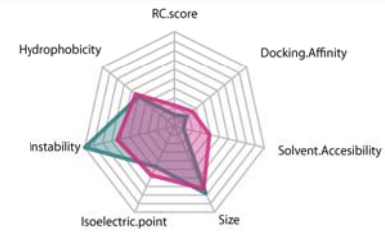
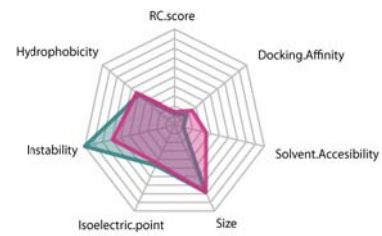
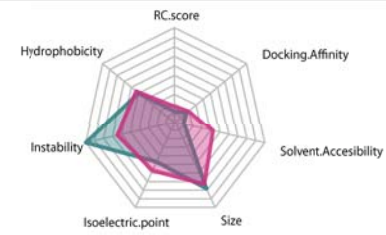
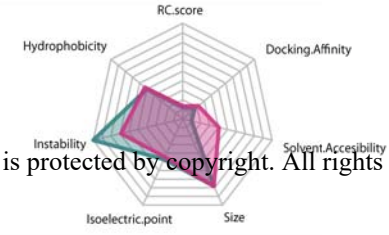
Binding to targets

8



Bound Luminescence

This article is protected by copyright. All rights reserved.

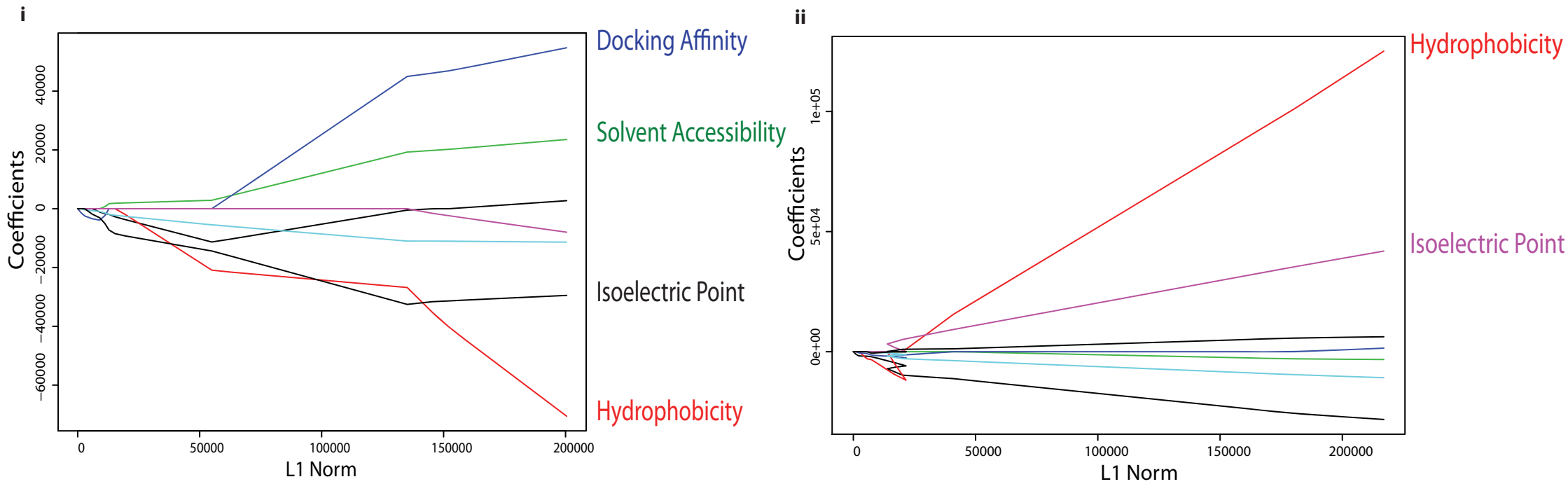
Test Sequence	F2F Result	F2F Score
ClfA Diabody 1		6.88
ClfA Diabody 2		9.48
ClfA Diabody 3		6.03
ClfA Diabody 6		9.27
ClfA Monobody 1		5.62
ClfA Monobody 2		3.62
ClfA Monobody 3		5.29
ClfA Monobody 4		3.8

Accepted Article

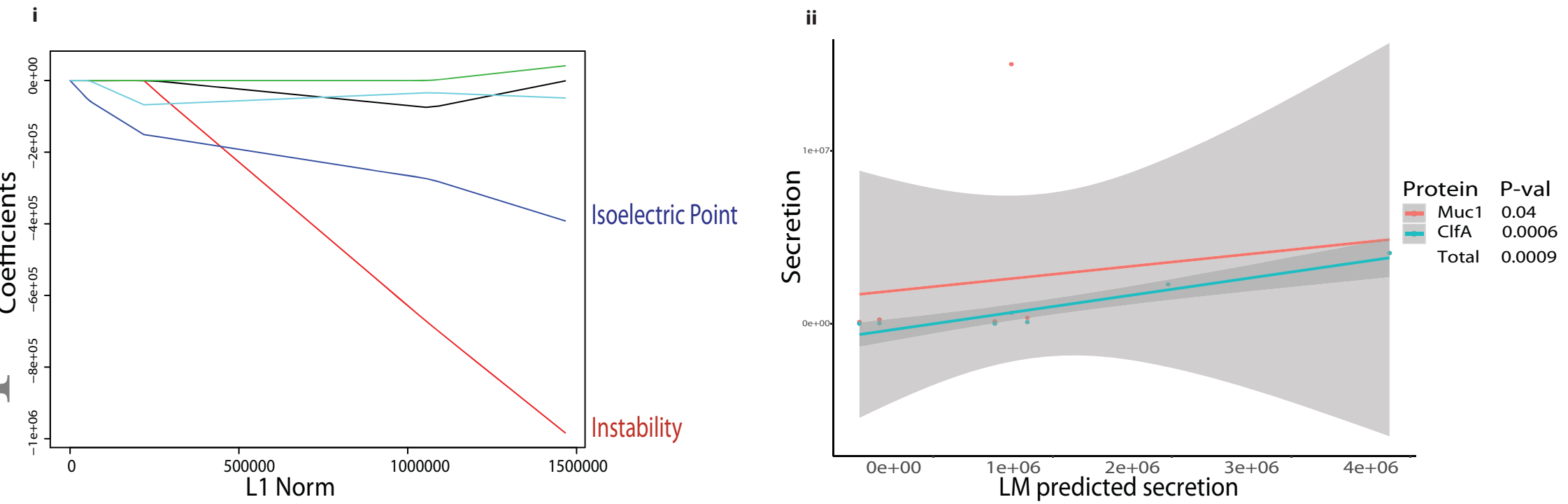
This article is protected by copyright. All rights reserved.

Accepted Article

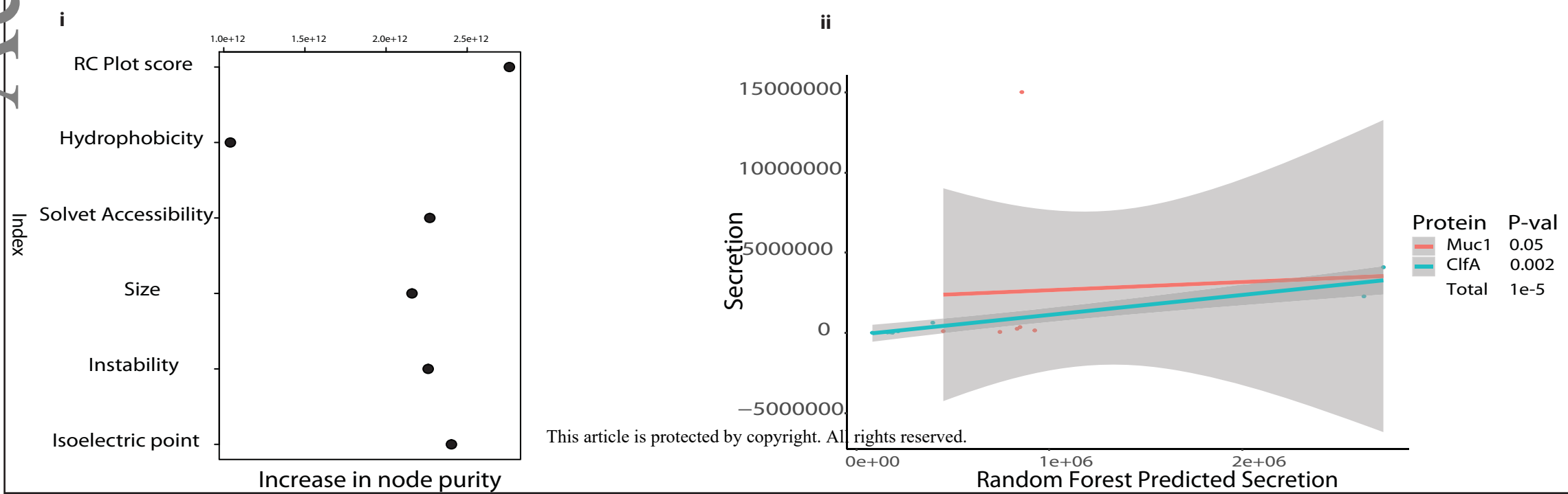
A



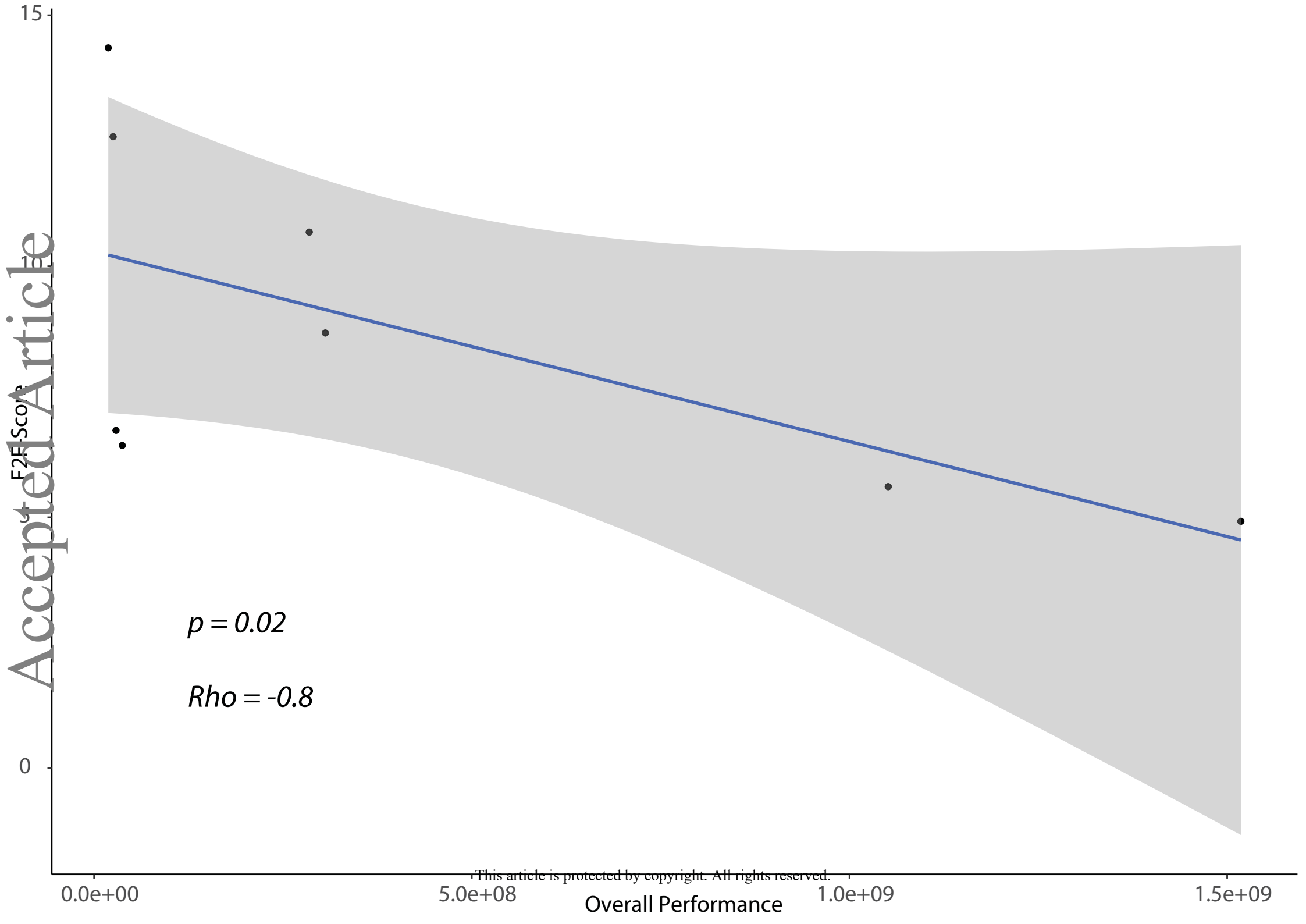
B



C



This article is protected by copyright. All rights reserved.



Accepted Article

This article is protected by copyright. All rights reserved.